A Random Forest-Based Machine Learning Framework with PCA, SMOTE, and SHAP for Efficient and Interpretable Coronary Artery Disease Prediction

T. Aswani[1*], Dr. Jose Moses Gummadi[2], Dr.G. Sharada[3]
[1]Research Scholar, Department of Computer Science and Engineering, School of Engineering, Malla Reddy University, Hyderabad, Telangana, India
[2]Department of Computer Science and Engineering, School of Engineering, Malla Reddy University, Hyderabad, Telangana, India
[3]Department of Computer Science and Engineering, Malla Reddy College of Engineering and Technology(A), Hyderabad, Telangana, India
Email: 2232CS010021@mallareddyuniversity.ac.in, josemoses@gmail.com, gsharada8@gmail.com
[*]Corresponding author

Keywords: machine learning framework, coronary artery disease prediction**,** feature engineering**,** class imbalance handling**,** model interpretability

Received: January 8, 2025

Given that coronary artery disease (CAD) is a major global cause of morbidity and mortality, there is an urgent need for precise and scalable diagnostic tools. While conventional machine learning (ML) models such as XGBoost and Gradient Boosting have demonstrated good predictive performance, they suffer from limitations, including weak handling of class imbalance, redundant feature spaces, and lack of interpretability. This work proposes an optimized Random Forest-based framework for CAD prediction to address these gaps, integrating advanced feature engineering and optimization techniques. Specifically, dimensionality reduction is achieved using principal component analysis (PCA), class imbalance is handled through the Synthetic Minority Oversampling Technique (SMOTE), and hyperparameter optimization is performed via GridSearchCV, tuning parameters such as the number of estimators, maximum depth, and minimum samples split. Additionally, SHAP (Shapley Additive exPlanations) values enhance interpretability by illustrating the contribution of each feature to the model's predictions; for example, features such as chest pain type and cholesterol level are shown to influence CAD outcomes significantly. The proposed framework is evaluated on the UCI Heart Disease dataset comprising 303 samples. Experimental results demonstrate that the optimized Random Forest model achieves an accuracy of 95.0%, outperforming Gradient Boosting (93.08%) and XGBoost (92.4%) classifiers. This framework provides a clinically relevant, interpretable, and scalable solution for CAD prediction, bridging the gap between technical advancements and their practical deployment in healthcare environments.

Povzetek: Razvit je izboljšan okvir za napovedovanje koronarnih arterijskih bolezni, ki uporablja algoritem naključnih gozdov, PCA za zmanjšanje dimenzionalnosti, SMOTE za ravnotežje razredov ter analizo SHAP za povečanje interpretabilnosti modela, kar omogoča klinično relevantno napovedovanje.

1 Introduction

Since coronary artery disease (CAD) is a primary global source of morbidity and death, early and precise diagnostic methods are crucial. Recent advances in machine learning (ML) have influenced CAD prediction, providing an excellent option to integrate clinical, diagnostic, and imaging data. Using a range of models, such as Gradient Boost and XGBoost, has shown promising results in predictive performance in existing studies. However, such methods have severe issues, such as class imbalance, redundant features, and inadequate generalizability. Moreover, the lack of interpretability inherent in most state-of-the-art approaches impedes their uptake in clinical practice.

These issues have been the subject of recent studies. For instance, Gupta et al. [13] utilized SMOTE and augmented features for high accuracy. However, Hashemi et al. [15] proposed integrating genetic algorithms with the one-layer multi-layer perceptrons for better predictions. While these approaches have been promising, they still leave a massive gap in scaling traditional ML models that achieve the optimal trade-off between accuracy, scalability, and interpretability. This work strives to address this gap by building upon existing implementations of CAD prediction to provide an optimized framework using RForest with improved feature engineering and advanced hyperparameter tuning methods.

This study attempts to create a machine-learning framework that enhances CAD prediction by overcoming the main limitations of existing approaches. The proposed methodology comprises a combination of PCA for dimensionality reduction, SMOTE for addressing the class imbalance, and GridSearchCV for hyperparameter tuning, among other novelties, leading to improved predictive performance and robustness. SHAP values are used for interpretability to elucidate feature contribution and improve model clinical relevance.

This study aims to develop a framework based on machine learning (ML) that includes three fundamental challenges in coronary artery disease (CAD) prediction: (1) clinical datasets suffer from class imbalance, (2) redundancy of features may lead to overfitting and generalizability, and (3) the lack of interpretability of the model. More specifically, we propose that the inclusion of PCA for dimensionality reduction, SMOTE for balancing classes, and SHAP values for interpretability of features used in the model, in conjunction with hyperparameter optimization Random Forest models, will provide better overall prediction accuracy robustness, and relevant to clinical practice than other pre-existing models such as Grad Boost and XGboost.

Specifically, the main contributions of this paper comprise: (1) an optimized implementation of Random Forest for CAD prediction; (2) the enhanced use of feature engineering techniques to boost model quality; (3) ensuring better interpretability of the model, using SHAP values, which is one of the significant drawbacks of currently used ML-based models; and (4) comparison against state-of-the-art models to validate our approach. The paper is organized as follows. In Section 2, existing CAD prediction methods are discussed, and research gaps are defined through a complete literature review. Section 3 outlines the proposed approach involving data preparation steps, feature engineering, and model tuning methods. Section 4 shows the experimental findings and assesses how well the suggested framework compared to state-of-the-art models. Section 5 presents the study's findings, contributions, and limitations, with Section 5.1 devoted to constraints. Finally, Section 6 summarizes the paper and discusses the study's overall importance for improving CAD diagnosis and patient care and possible future directions.

2 Related work

This literature review highlights advancements in machine learning-based approaches for coronary artery disease (CAD) prediction and management. Bertsimas et al. [1] developed ML4CAD, a personalized CAD management system with an 81.5% AUC, using EMR data. Future work will include validating clinical trials, including socioeconomic characteristics, and improving generalizability. Varuna et al. [2] recommended applying a two-phase AI model to identify coronary artery disease with 96.2% accuracy. More studies will try to make it more generalizable and expand its use to include other illnesses. Gabriel Anbarasi [3], the BSOXGB model outperforms previous approaches with a CAD recognition success rate of 97.70% because of enhanced feature selection and hyperparameters. Testing with additional datasets will be part of future efforts. Huang et al. [4], the RF model predicts CHD remarkably effectively using CACS and clinical factors. Future work will include handling missing data and improving models. Manduchi et al. [5] demonstrated how well TPOT can detect SNPs linked to CAD, while it has issues with big datasets, runtime, and heterogeneous data.

Zahia et al. [6] used feature selection and data balance to develop a hybrid machine-learning model that detects CAD with 98.34% accuracy. Jahmunah et al. [7] introduced a GaborCNN model for ECG-based CAD diagnosis with a 98.5% accuracy rate and potential for faster, more effective clinical use. Arian et al. [8] predicted myocardial function improvement after CABG using LGE-CMR images radiomics and machine learning, with encouraging findings. Umar Khan et al. [9] suggested a signal processing technique for CAD prediction that uses ECG data and SVM. It achieves 95.5% accuracy and suggests deep learning for future advancements. Nasarian et al. [10], a hybrid feature selection approach for CAD, is presented in this work, which achieves excellent accuracy by utilizing a variety of classifiers and balancing strategies. Expanding datasets and investigating evolutionary algorithms are the goals of future research.

Abdar et al. [11] proposed a novel machine-learning approach for CAD identification with an accuracy of 93.08%. Additional preprocessing methods, algorithms, and evolutionary approaches will be investigated in further studies. Li et al. [12] improved risk group categorization by creating a framework for risk stratification with machine learning assistance to streamline CAD diagnosis. Ongoing research might develop these techniques further. Gupta et al. [13] said that the C-CADZ system outperformed earlier techniques in achieving 97.37% accuracy for CAD diagnosis utilizing FAMD and SMOTE. Future research might improve multi-class categorization and the handling of class imbalance. Varun et al. [14], a deep neural network diagnosed CAD with 96.2% accuracy using Gaussian noise to reduce overfitting; future work will concentrate on extending to other ailments. Hashemi et al. [15] employed genetic algorithms and machine learning to predict CAD with 94.71% accuracy; deep learning advancements will be the main focus of future work.

Nesaragi et al. [16] presented a tensor-based machine learning system that achieves 96.62% accuracy in CAD identification using heart rate data. Further research will improve this approach. Saruladha and Swathy [17] examined AI and data mining strategies for predicting CVD, emphasizing the need for more information and customized approaches. Huang et al. [18] AI accelerates productivity and improves the accuracy and efficiency of computed tomography angiography (CCTA), a technique used to diagnose computer-aided design (CAD). Khozeimeh et al. [19] Active Learning with Ensemble of Classifiers (ALEC) improves the diagnosis of CAD by lowering the risks and costs associated with invasive angiography. Qiao et al. [20] suggested that ML-based FFRCT may improve CAD diagnosis and decision-making compared to invasive angiography; nevertheless, more validation is needed.

Alizadehsani et al. [21] presented a high-accuracy machine-learning approach to identifying individual cases of coronary artery stenosis while resolving model uncertainty. Omkari and Shaik's [22] TLV model uses ensemble voting and machine learning to achieve high accuracy in CAD diagnosis on large datasets, which makes it perform better than previous methods. Braun et al. [23], a non-invasive, precise, and economical CAD screening technique, is provided by cardiography, which combines vector cardiography with machine learning. Wishart et al. [24] provided a cost-aware feature selection technique to identify coronary heart disease with excellent accuracy and AUC using fewer features. Wang et al. [25] provided a two-level stacking machine learning model for CHD diagnosis. Although it produces high accuracy, the dataset's amount and the parameters' values are limited.

Ahmad et al. [26] LR, KNN, SVM, and GBC approaches are examined in this study and shown to be less accurate than Extreme Gradient Boosting with GridSearchCV in predicting cardiac disease. Yan et al. [27] created a machine learning-based system that uses XGBoost for high classification accuracy and customized patient advice to diagnose coronary artery stenosis. Cheung et al. [28] provided a 2D UNET model for accurately segmenting coronary arteries on CTCA photos using the least computer resources. Spadarella et al. [29] Radiomics and machine learning deliver promising advancements to cardiovascular imaging despite ongoing issues with study consistency and model interpretability. Benjamins et al. [30] state that the capacity to identify myocardial ischemia and the necessity for early revascularization is improved by combining machine learning with CTA and clinical data, albeit further validation is needed.

Molenaar et al. [31] Artificial intelligence in invasive coronary angiography is advancing, even if more multicenter datasets and external validation are required for broader applications. Muhammad and Algehyne [32] enhanced the C4.5 algorithm, which was used to create a fuzzy-based expert system for CAD in Nigeria that produced high dependability and accuracy of 94.55%. Hagan et al. [33] highlighted different training costs and accuracy across datasets when comparing machine learning techniques for diagnosing cardiovascular disease. Brandt et al. [34] assessed CT-derived fractional flow reserve (CT-FFR), which may reduce the requirement for invasive angiography to identify substantial CAD in patients with severe aortic stenosis. Liu et al. [35] evaluated a machine-learning model for severe CAD prediction using routine data to reduce invasive procedures and improve diagnosis accuracy.

Table 1: Summary of existing machine learning models for CAD prediction

Study (Author, Year)	**Dataset Used**	**Methodology / Model**	**Performance (Accuracy)**	**Interpretability Addressed**	**Class Imbalance Handling**	**Dimensionality Reduction**
Bertsimas et al. (2020) [1]	EMR Data	ML4CAD (Multiple Models)	81.5% (AUC)	No	Not explicitly addressed	Not addressed
Varuna et al. (2023) [2]	Custom Dataset	Two-phase AI Model (Deep Learning)	96.2%	No	Not explicitly addressed	Not addressed
Gabriel et al. (2023) [3]	Public Dataset	BSOXGB (XGBoost + Feature Selection)	97.7%	Partial (XGBoost + SHAP support but not emphasized)	Not mentioned	Feature Selection (not PCA)
Zahia et al. (2020) [6]	Clinical Dataset	Hybrid ML Model with Feature Selection	98.34%	No	Balancing techniques used	Feature Selection (not PCA)
Jahmunah et al. (2021) [7]	ECG Data	GaborCNN (Deep Learning)	98.5%	No	Not addressed	Not addressed
Abdar et al. (2019) [11]	UCI Cleveland Dataset	Hybrid ML Approach	93.08%	No	Not specified	Not specified

Gupta et al. (2021) [13]	Z-Alizadeh Sani Dataset	C-CADZ (FAMD + SMOTE + Classifier)	97.37%	No	SMOTE used	Feature Aggregation (FAMD)
Hashemi et al. (2024) [15]	Public Dataset	Genetic Algorithm + Optimized MLP	94.71%	No	Not mentioned	Genetic Algorithm (not PCA)
Wang et al. (2020) [25]	Public Dataset	Stacking Ensemble Model	90.0%	No	Not specified	Not addressed
Benjamins et al. (2021) [30]	Clinical + CTA Data	XGBoost	92.4%	Not focused	Not specified	Not addressed
Proposed Study (2024)	UCI Heart Disease Dataset (303 samples)	Optimized Random Forest + PCA + SMOTE + SHAP	95.0%	Yes (SHAP used explicitly)	SMOTE applied	PCA applied (8 components)

Militello et al. [36] showed that integrating radiomic and clinical variables enhances the prediction of coronary artery disease compared to utilizing clinical data alone. Nilashi et al. [37] demonstrated how incremental machine learning techniques, especially fuzzy support vector machines (SVM), improve the precision of heart disease detection while cutting down on computation time. Raparelli et al. [38] integrated a variety of characteristics and used machine learning to separate obstructive from non-obstructive CAD; nevertheless, more enormous datasets are required for validation. Yang et al. [39] utilized enhanced LightGBM and focal loss, and the HY_OptGBM model improved early CHD diagnosis with a 97.8% AUC. Cherradi et al. [40] suggested that KNN and ANN-based diagnostic systems outperformed earlier techniques with higher accuracy for predicting atherosclerosis. Significant learning and data mining techniques have obtained up to 60% to 90% prediction accuracy by selecting features through an effective model and addressing the class imbalance problem. Sadri Alija et al. [42] used a supervised learning model and a wrapper-based feature selection component to improve student performance prediction over imbalanced datasets. Harjinder Kaur et al. [43] proposed a prediction framework based on academic performance analysis using machine learning algorithms focusing on the early detection of underperformers. Hua Huang [44] proposed a two-stage feature selection method and enhanced machine learning classifiers for text data classification. These studies demonstrate that optimized feature selection and balanced data learning are critical elements of predictive modeling. A comparative summary of key machine learning studies on CAD prediction is provided in Table 1. Other features include datasets, methodologies, performance, interpretability focus, class imbalance handling, and dimensionality reduction techniques. It also shows the novelty of the proposed approach while addressing the limitations of other existing works and providing robustness and clinical applicability through PCA, SMOTE, hyperparameter tuning, and SHAP interpretability as the final methodologies.

The studies surveyed revealed a range of machine learning models employed for CAD prediction, such as Random Forest, SVM, and hybrid models, resulting in notable accuracy. Feature selection, class balancing, and deep learning — all of these techniques lead to better performance. In future studies, we aim to expand the generalizability, work with larger datasets, as well as incorporate clinical, radiomic, and socioeconomic variables to make the CAD diagnosis robust.

3 Proposed framework

Figure 1 overviews the methodology for predicting coronary artery disease, which represents a systematic process adopted to implement a robust and accurate machine-learning framework. In the first step, the raw dataset was preprocessed. Feature Scales were standardized using the StandardScaler to normalize the data so that each feature contributed equally during the model's training. PCA was applied to the data, so the dimensionality of the data cube was reduced to eight principal components, which provide the best data features without raising the computation price and the possibility of overfitting.

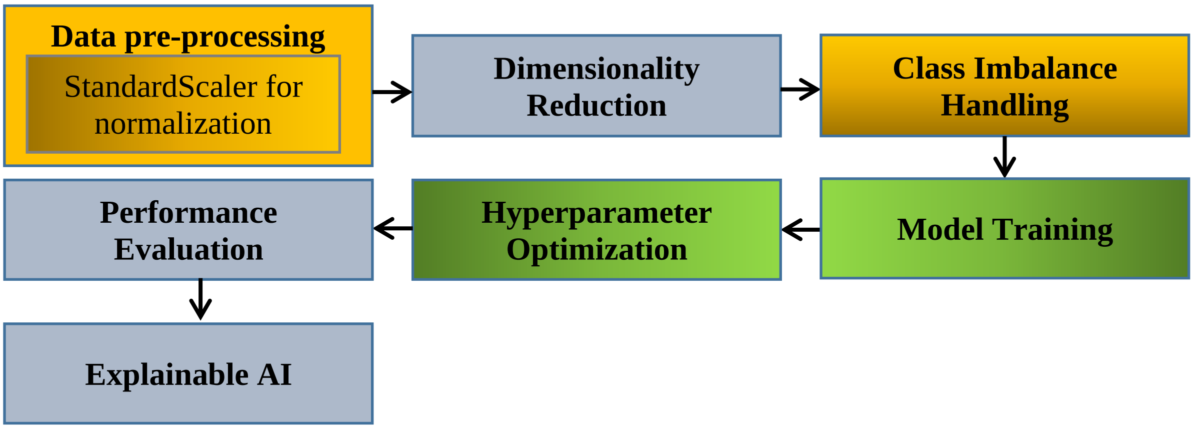

Figure 1: Proposed methodology for coronary artery disease prediction

To rebalance the target, we used the Synthetic Minority Oversampling Technique (SMOTE). Variable. Using this technique, synthetic samples were created for the minority class, which provided a balanced dataset for the model to learn patterns for both classes more effectively. The cleaned and balanced dataset was subsequently chosen for its robustness and appropriateness for high-dimensional data analysis and utilized to train a random forest classifier. Hyperparameter optimization was performed with GridSearchCV by searching over the Using a 5-fold cross-validation technique, the optimized model produced dependable and generalizable findings for the number of estimators, the maximum tree depth, and the least number of samples required to divide a node.

After that, the dataset was split into a 70:30 training-testing set, with the testing set used to assess the model's performance and the training set used to build the model. The performance metrics were based on accuracy, confusion matrix, and classification report (precision, recall, and F1-score). The SHAP (SHapley Additive exPlanations) framework was employed to improve model interpretability. SHAP values calculated the contribution of each feature to predictions, and a summary plot graphically described feature importance, revealing insights on the key predictors of coronary artery disease. The best model was serialized using joblib and exported to a. pkl file, making it suitable for clinical decision support systems deployment. This pipeline, outlined in Figure 1, integrates robust preprocessing techniques complemented by class balancing, hyperparameter tuning, and interpretability to create a strong and transparent framework for practical deployment.

3.1 Machine learning models

The performance of several machine learning models was compared in this study to predict coronary artery disease. Using the KNN technique, because it is a simple yet effective instance-based learning approach, we use the majority class observed among their closest neighbors to classify similar data points. This algorithm can find various local patterns which are the most relevant to the dataset. Furthermore, The SVM's capacity to identify the best hyperplane for dividing data points into distinct classes was also utilized. SVM is beneficial for high-dimensional data and creating robust decision boundaries. The Decision Tree Classifier was also one of the models evaluated (selected due to interpretability and simplicity). That's why this algorithm makes a tree structure that splits the data repeatedly using the value of a particular feature. We also used random Forest, a type of ensemble learning that builds numerous decision trees to increase precision and decrease overfitting. Random Feature Selection and Bootstrap Methodology The above methodology is inherent in Random Forest and is used to contribute to the robustness and reliability of the model partially. These various models gave the study a comprehensive assessment of different classification techniques. The performance of all models was analyzed and compared to the data for identification of coronary artery disease.

3.2 Data preprocessing

Data was preprocessed to prepare for machine learning modeling. We used the StandardScaler to standardize each feature, assigning a standard deviation of one and a mean of zero to the data. This resolution ensured that broader-scale features did not disproportionately affect the model. No null values remained. They were one-hot encoded to prepare categorical variables for use in K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Trees, and Random Forest models. This preprocessing step was crucial so that all models performed consistently.

First, the dataset was preprocessed to handle missing and categorical data encoding. In particular, categorical variables (e.g., types of chest pain and thalassemia) were preprocessed using one-hot encoding to convert them to numeric format. Feature scaling was executed with StandardScaler to have all features with a zero-mean and one-variance distribution. MinMaxScaler, RobustScaler, and other scaling techniques were tried but again showed a minuscule impact on performance, so the best choice was to use StandardScaler, especially before PCA.

3.3 Dimensionality reduction

Feature dimensionality was reduced using Principal Component Analysis (PCA) after the data were scaled, retaining 95% of the data variance. This threshold was empirically derived; retention of 90% variance was found to exclude clinically relevant features dangerously, and retention of 99% yielded trivial variance benefit with increased model complexity. By eliminating duplicate and correlated features, PCA helped to prevent overfitting and enhanced the efficiency of our models. While PCA is a linear combination of features and thus could impact direct clinical interpretability, the SHAP values were calculated on features before PCA was performed, which ensures the interpretation of feature contributions.

3.4 Handling class imbalance

Instead of a class imbalance (majority class), the target variable showed a high-class imbalance. For the minority classes, synthetic samples were made using the Synthetic Minority Oversampling Technique (SMOTE). This resulted in a balanced distribution required for all four machine learning models to be trained relatively and practically. SMOTE stopped SVM and KNN, Decision Tree, and Random Forest models from learning patterns only on majority classes, thus enhancing the accuracy of their predictions.

SMOTE handled imbalance classes after the one-hot encoding and feature scaling process but before applying PCA transformation. The dataset was also imbalanced, with 55% majority (non-CAD) and 45% minority (CAD) samples. The SMOTE generated synthetic samples for the minority class, balancing the classes' ratios to 50:50, ensuring that the classifier learned the same amount from both classes and improving recall and F1-score.

3.5 Model training

The preprocessed and balanced dataset trained four models (i.e., KNN, SVM, Decision Tree, and Random Forest). On the contrary, KNN predicted a class of data points using a majority vote among the nearest neighbors, so it followed the local pattern of the data. This is due to SVM's capability to work with high-dimensional data by identifying an optimal hyperplane to distinguish between. The Decision Tree model, which segmented the data based on feature values, provided an interpretable structure for decision-making. The Random Forest was an ensemble of decision trees trained using bootstrapping and random feature selection to reduce variance while increasing accuracy. Each of these models was trained separately to provide a complete evaluation.

An A 5-fold cross-validation technique was incorporated during hyperparameter tuning and model evaluation to reduce the chance of overfitting. The dataset was randomly divided into a training set (70% of the data) and a testing set (30% of the data), and all works used a fixed random seed value of 42 to guarantee the replicability of the results. Using a confusion matrix, we used professional metrics to assess models, including accuracy, precision, recall, F1-value, and analysis. To account for class imbalance in the predictive model, cross-validation was used to select the optimal parameters for the model by maximizing the F1 score, which balances performance in both classes. The training set was utilized for model construction and hyperparameter tuning. The testing set measured the model performance using data never used during training. The testing set can thus be viewed as providing an unbiased performance result for a particular application. Performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis were evaluated. GridSearchCV yielded an exhaustive hyperparameter search, but RandomSearchCV and Bayesian optimization could have also been applied. Due to the moderate size of the dataset with a well-defined parameter space, GridSearchCV was favored for its systematic search strategy without introducing an excessive computational overhead.

Initial experiments included KNN to establish baseline performance, leveraging its ability to capture local data patterns effectively. However, the final optimized model employed Random Forest, chosen for its robustness, superior generalization, and ability to handle feature interactions, which proved essential for CAD prediction.

3.6 Hyperparameter optimization

This way, hyperparameter optimization with GridSearchCV was carried out to obtain each model's peak performance. KNN: suitable number of neighbors and SVM: tunning parameters like kernel and regularization strength. The maximum depth and split criteria were fine-tuned for the Decision Tree model. The number of estimators, maximum depth, and minimum samples for the split were optimized for Random Forest. This led to a 5-fold cross-validation, in which the optimized parameters were replicated across all models, resulting in accurate and generalizable results.

The hyperparameter tuning of the Random Forest model was applied using GridSearchCV with a 5-fold cross-validation strategy. The explored range of parameter values was: number of estimators [50, 100, 200, 300], maximum depth [4, 6, 8, 10, None], minimum samples split [2, 5, 10], and the minimum samples leaf [1, 2, 4]. A random seed of 42 guaranteed that query results could be reproduced. By restricting maximal depth and tuning minimum examples of leaves, Random forests, in their nature, enforced regularization, which kept away overfitting. The F1-score has been in the first place in choosing the best hyper-parameters because we have imbalanced classes; we need to balance precisely and recall.

3.7 Explainable AI (SHAP)

We implemented the SHAP (SHapley Additive exPlanations) framework to interpret the predictions from ML models. The significance of each attribute in predictions was evaluated using SHAP values, which describe the reasoning of KNN, SVM, Decision Tree, and Random Forest. Global interpretation plots (summary plots) were created to examine the relative contributions of features across all the models. This also improved transparency and made the models' predictions explainable, allowing them to be used in a clinical setting, where it is essential to know why a decision is made.

3.8 Proposed algorithm

The proposed Intelligent Coronary Artery Disease Prediction (ICADP) algorithm uses four optimized Random Forest, SVM, KNN, and Decision Tree machine learning models. It combines advanced preprocessing, class balancing, and interpretable predictions via SHAP. Without loss of generality, this algorithm generates a robust, fair, and interpretable predictor that is highly meaningful in clinical settings toward promoting informed healthcare decision-making.

Algorithm: Intelligent Coronary Artery Disease Prediction (ICADP)
Input: Dataset (X, Y), Models M = { Random Forest, SVM, Decision Tree, and KNN}, Parameters P_m for m in M
Output: Optimized Models M*, Metrics for each model, Predictions Y_pred

```
1. Preprocess X: Normalize features, handle missing values
2. Reduce Dimensions:
   X_PCA ← PCA(X, retain 95% variance)
3. Handle Class Imbalance:
   (X_balanced, Y_balanced) ← SMOTE(X_PCA, Y)
4. Split Data:
   (X_train, Y_train), (X_test, Y_test) ← Train-Test-Split(X_balanced, Y_balanced)
5. Train and Optimize Models:
   For each model m in M:
      m* ← GridSearchCV(m, P_m, cv=5)
      m*.fit(X_train, Y_train)
6. Evaluate Models:
   For each optimized model m* in M*:
      Y_pred_m ← m*.predict(X_test)
      Metrics_m ← Evaluate(Y_test, Y_pred_m)
7. Interpret Results:
   For each m* in M*:
      SHAP_values_m ← SHAP(m*, X_test)
8. Return M*, Metrics_m, Y_pred_m, SHAP_values_m
```

Algorithm 1: Intelligent coronary artery disease prediction (ICADP)

For accurate profiling of coronary artery disease, the ICADP algorithm systematically utilizes various ML models to implement and predict CAD effectively. The next step is to preprocess the dataset to fit it into the machine learning format. The features are standardized with StandardScaler, which gives them a mean of zero and a standard deviation of one. So, this step removes bias from different scales among the features to ensure consistency. It also handles missing values and encodes categorical variables to guarantee that the information aligns with the models. A PCA is applied to this initial dataset to reduce dimensionality. Retaining only the eight principal components with the highest variance helps the algorithm retain significant data while discarding redundancy, streamlining the feature space, and reducing the likelihood of overfitting. The Synthetic Minority Oversampling Technique balances the dataset's classes. To combat this, SMOTE creates artificial samples for the minority class and, in turn, gets a better-balanced dataset to ensure that all models learn better for both class patterns. This preprocessed, balanced data is split into training and testing different subsets, with a ratio of 70 to 30 to guarantee sufficient data for model evaluation and training.

Next, four machine learning models were used: Random Forest, Decision Tree, Support Vector Machine, and K-Nearest Neighbors (KNN). These are optimized individually through hyperparameter tuning with the help of GridSearchCV. We are running hyperparameter tuning with a different grid search space for each model for tuning settings like the number of KNN neighbors, SVM kernel type, Decision Tree maximum depth, and Random Forest number of estimators. The optimization employs a 5-fold cross-validation strategy, resulting in reliable and generalizable results for every model.

After these steps have been optimized, the selected models are trained with the balanced training data and evaluated with the testing dataset. Predictions are obtained from each model, and performance metrics such as F1-score, confusion matrix, recall, accuracy, and precision are calculated. These metrics give a complete analysis of the effectiveness of each model and a comparison of performance. The Shapley Additive Explanations, or SHAP framework, is used to improve the interpretability of the algorithm. Computing SHAP values abstracts how much each feature has contributed to the predictions and, thus, provides a comprehensive view of all four models' prediction logic. We generate summary plots to visualize how features rank globally in importance, establishing a transparent basis to support transferable clinical uses of our interpretable framework. Ultimately, the ICADP algorithm produces the tuned models, those model's performance metrics, and the SHAP-based interpretations. This end-to-end process also guarantees robustness, accuracy, and interpretability for the employed, leading to generalizability that renders the framework well-suited for real-world coronary artery disease prediction scenarios.

3.9 Dataset details

The dataset [41] that is used to predict coronary artery disease consists of 303 samples with 14 attributes, i.e., Age, Sex, Chest Pain Type (cp), Resting Blood Pressure (trtbps), Cholesterol Level (chol), Fasting Blood Sugar (FBS), Resting Electrocardiographic Result (restecg), Maximum Heart Rate Achieved (thalachh), Exercise Induced Angina (exng), ST Depression Induced by Exercise (oldpeak), Slope of Peak Exercise ST Segment (slp), No of Major Vessels (caa) and Thalassemia (thall). The target variable (output) is a binary value indicating whether the patient has coronary artery disease. This dataset contains a rich feature set of clinical and demographic characteristics that can aid in building and testing machine learning models.

3.10 Performance evaluation

The performance of each model was evaluated using f-score, recall, accuracy, precision, and confusion matrix. Although accuracy offered an overall assessment of correctness, precision and recall allowed us to assess the models' performance in every class. The confusion matrix provided rich detail on true positives, negatives, false positives, and false negatives. The desired classifier that performed best for the out-of-sample was identified using these metrics, and we compared the KNN, SVM, Decision Tree, and Random Forest models.

3.11 Experimental setup

All experiments are done on Python 3.9 with the sci-kit-learn 1.2.2 library for ML models. Other libraries employed were pandas 1.4.3, numpy 1.21.5, and matplotlib 3.5.2 to assist in the data processing and visualization. What exists needs to be replaced by what better exists (or, in other words, by data that is a better approximation). The experiments were done on an Intel Core i7-12700 CPU,16GB RAM, and Windows 11 OS. The random seed value was set to 42 for all runs to guarantee reproducibility. GridSearchCV was used for hyperparameter tuning with 5-fold cross-validation, applying the same search space to the key parameters of each classifier. Also, we trained and evaluated our models in a non-parallelized manner to keep the computational conditions consistent for all models.

4 Experimental results

Results from the experiment are shown in this section. Creating a coronary artery disease prediction strategy is the goal of the suggested system using an extensive clinical and diagnostic data dataset. To benchmark the proposed framework, comparative experimentation is performed against state-of-the-art machine learning models such as Gradient Boosting [26], XGBoost [30], and Logistic Regression [26]. We conducted these experiments in Python with sci-kit-learn and other libraries on a computer with 16GB of RAM, an Intel Core i7, and an NVIDIA GPU for speeding computations. The analysis examines the effect of feature engineering (PCA and SMOTE) and hyperparameter tuning on predictive accuracy and model robustness.

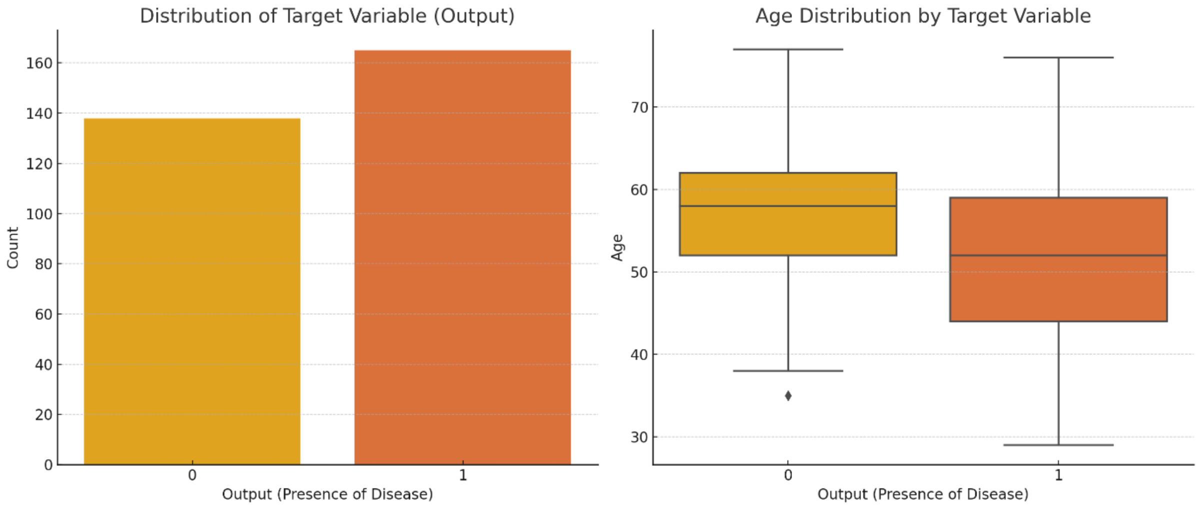

Figure 2: Distribution of the target variable (presence of disease) and age distribution by target variable for coronary artery disease prediction

The overview of the target variable (output) is illustrated in Figure 2, where we can observe that class 1 (disease is present) is slightly more frequent than class 0 (disease is absent). Boxplot of age shows that patients who have coronary artery disease (class 1) have a broader range of ages than those who don't (class 0). The median age of patients with CAD is also higher; thus, the age is a predictor for CAD. This visualization highlights the need for balanced class representation and age consideration during your model development.

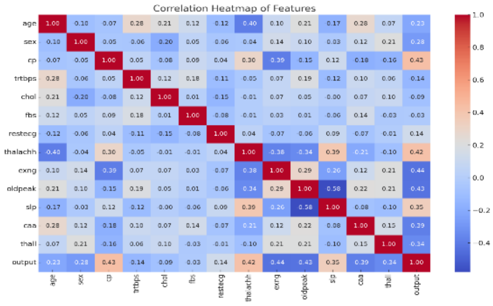

Figure 3: Correlation heatmap of features illustrating the relationships between variables and their influence on the target variable (output) for coronary artery disease prediction

The correlation heatmap, which represents the pair-wise relationships and the correlation of the data with the target variable (output), is shown in Figure 3. A strong positive correlation can be observed for cp(chest pain type), halacha (highest heart rate attained), and SLP (slope of peak exercise ST segment) with output, which signifies their importance in CAD prediction. Conversely, attributes such as exng (exercise-induced angina) and old peak (ST depression) exhibit strong negative correlations. The heatmap also shows that there isn't much multicollinearity amongst most features, confirming these to be good candidates for model training. This analysis provides significant predictors in machine learning models that aid feature selection and optimization.

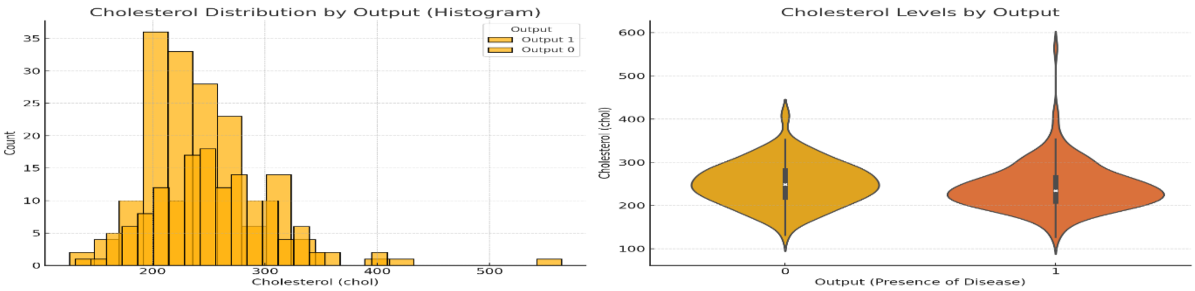

Figure 4: Cholesterol distribution by target variable (output) depicted using a histogram and a violin plot for coronary artery disease prediction

An example of using this method to visualize information about a categorical variable is to look at the stimulus across the target output (Figure 4). One of the classical diagnosis methods is to analyze cholesterol levels. A histogram indicates overlapping cholesterol levels for both classes (1, 1). There is a relatively higher concentration of samples in 200 and between 300. In particular, the violin plot elucidates the spread and density of the cholesterol levels, suggesting higher median cholesterol values among patients with coronary artery disease (output = 1). Cholesterol variability across CAD patients is indicated by class 1 having a wider distribution. These findings emphasize cholesterol as a key attribute for CAD prediction, but additional features could improve class discrimination.

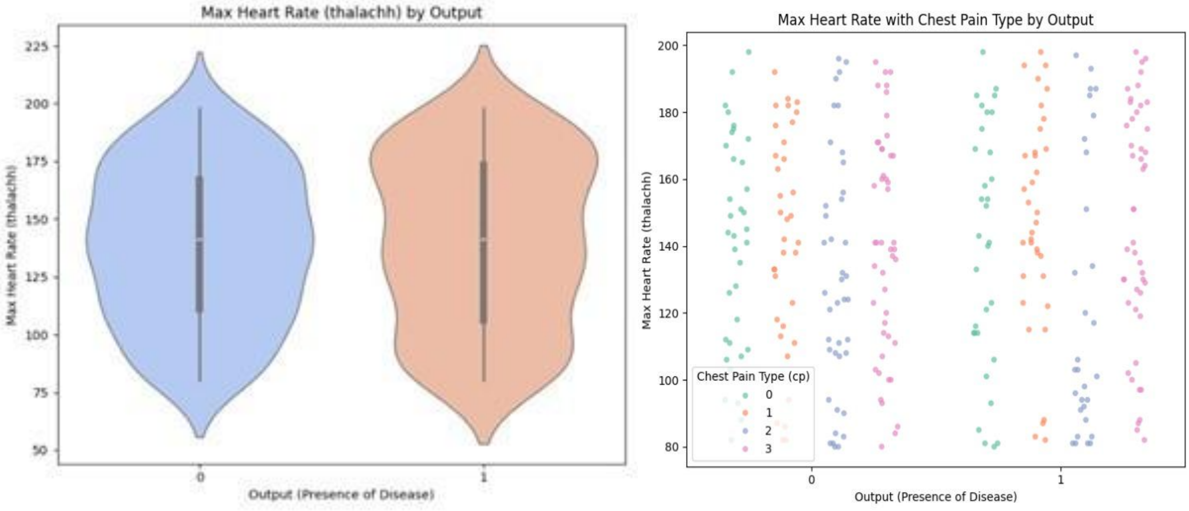

Figure 5: Max heart rate (thalachh) distribution by target variable (output) depicted using a violin plot and its relationship with chest pain type (CP) for coronary artery disease prediction

An example of a violin plot showing the maximum heart rate (thalachh) distribution by human-readable target (output) illustrating similar distributions for all CP levels can be found in Figure 5. The median heart rate unit (1) is higher in patients presenting with coronary artery disease, with a broader variance compared to another unit (0). This second plot adds chest pain types to the mix, showing how heart rate distributions by class differ. To detail this with some visualization and explain how this is an important identifying feature and Analysis of the interaction between heart rate and chest pain type provides essential information for CAD prediction models.

Figure 5 illustrates that patients with coronary artery disease (unit 1) tend to exhibit higher maximum heart rates with more significant variance compared to non-CAD patients (unit 0). The second plot shows that typical angina (cp=0) is associated with lower heart rates within the CAD group. In contrast, atypical chest pain types (cp=1,2) correspond to higher heart rates, highlighting distinct patterns relevant for clinical assessment.

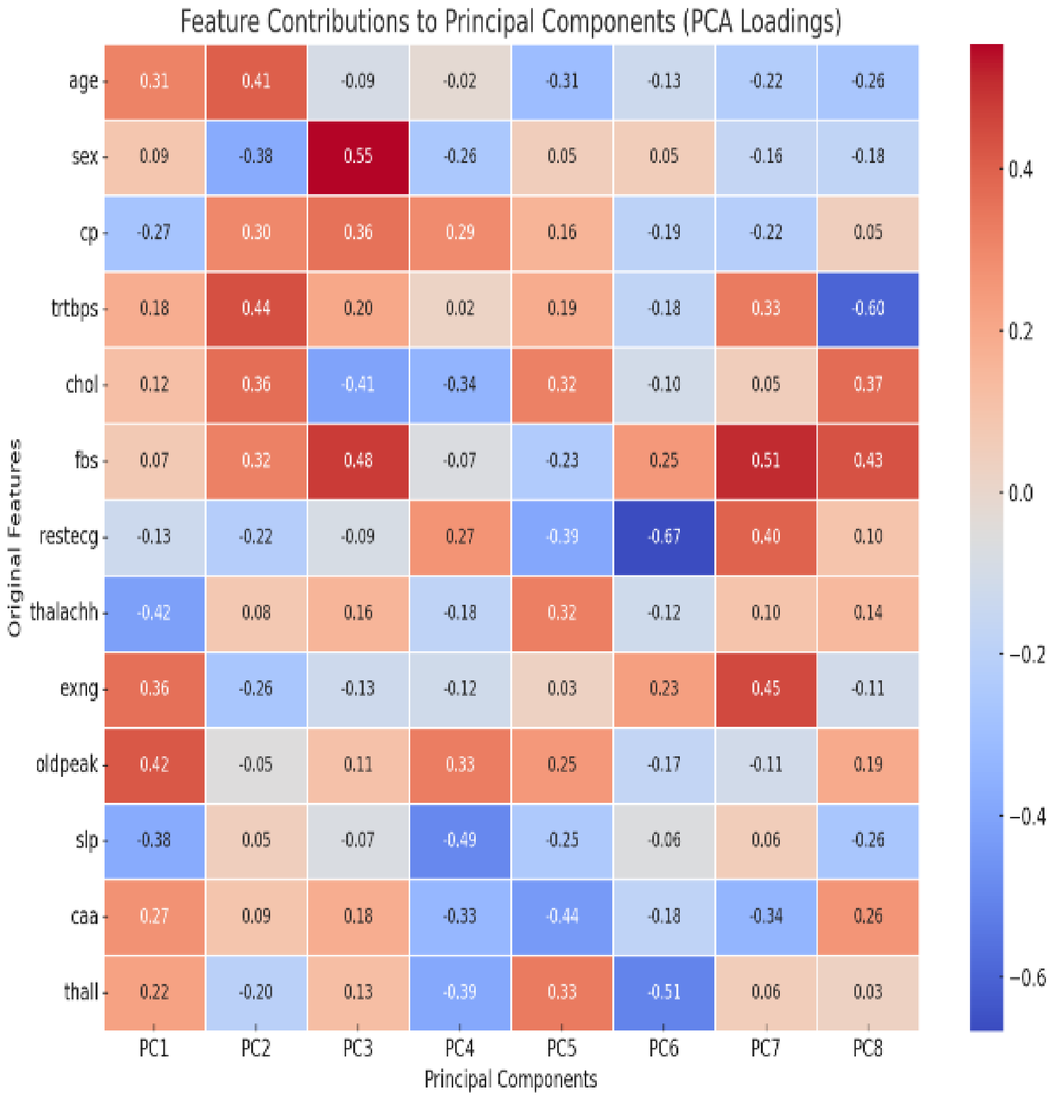

Figure 6: Feature contributions to principal components (PCA loadings) illustrating the relationship between original features and principal components for dimensionality reduction in coronary artery disease prediction

The PCA loading of original features that contribute to the eight principal components is depicted in Figure 6. Based on unsupervised feature selection analysis, it is clearly shown that features like cp (chest pain type), halacha (max heart rate), and old peak (st depression) have significant contributions in the first few components, meaning that these are essential features in capturing variance in the dataset. On the other hand, attributes such as resting (resting electrocardiographic results) are less influential across components. This exploration shows the capacity of dimensionality reduction with PCA implementation to build upon the dataset's most significant features to deliver the model's maximum performance while conserving important predictive knowledge content to be kept for coronary artery disease diagnosis.

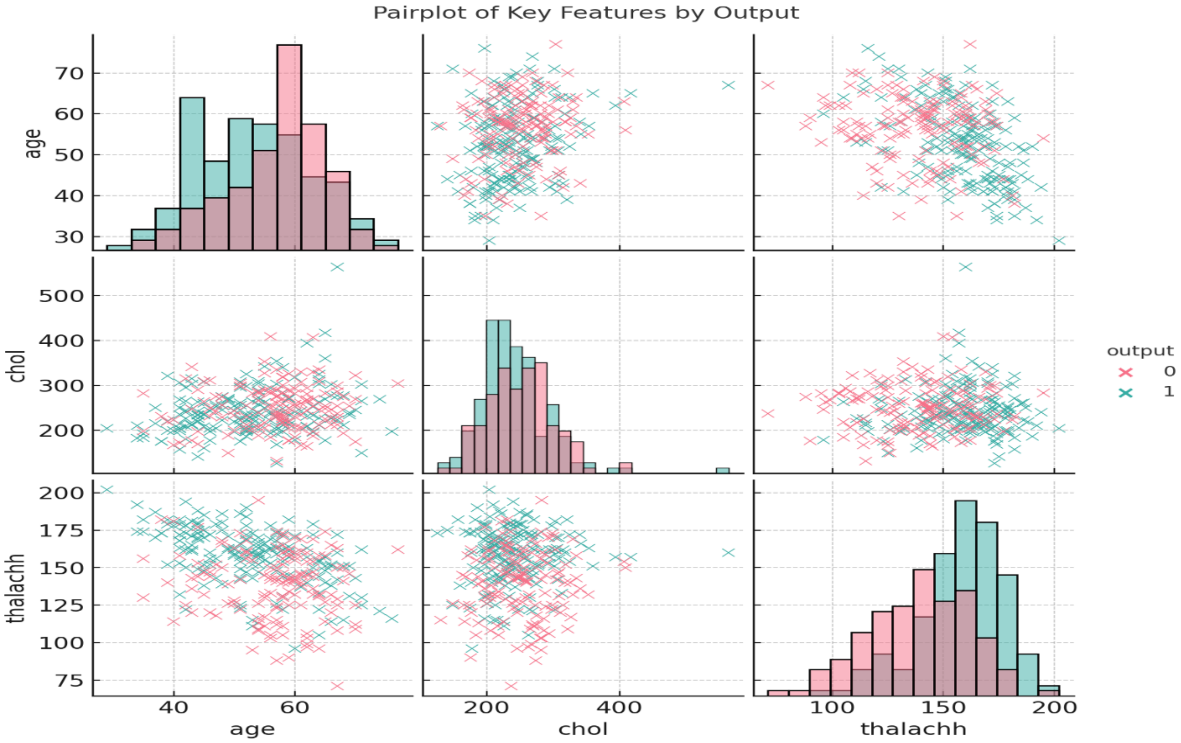

Figure 7: Pairplot of key features (age, chol, and thalachh) by target variable (output) for coronary artery disease prediction

We have plotted the pairplot that can correlate the primary features (age, chol, and thalachh) across the target variable (output) between each other (refer to figure 7). If we look at the diagonal histograms, we can see that most features overlap between the two classes. Still, we can also identify subtle differences (such as higher thalachh when CAD = 1). Scatterplots reveal weak correlations among features, reinforcing the relevance of those variables when paired in the context of ML models. The visualization can help with the separability and interaction of features that can be used to predict CAD.

Table 2: Hyperparameter tuning details for ML models, including the hyperparameters considered, their respective search spaces, and the optimized values

Model	**Hyperparameter**	**Hyperparameter Space**	**Optimized Value**
KNN	Number of Neighbors (k)	[3, 5, 7, 9, 11]	5
SVM	Kernel	['linear', 'rbf', 'poly']	'rbf'
	C (Regularization)	[0.1, 1, 10, 100]	10
	Gamma	['scale,' 'auto']	'scale'
Decision Tree	Maximum Depth	[5, 10, 15, None]	10
	Minimum Samples Split	[2, 5, 10]	5
	Criterion	['gini,' 'entropy']	'gini'
Random Forest	Number of Estimators	[50, 100, 150, 200]	150
	Maximum Depth	[5, 10, 15, None]	15
	Minimum Samples Split	[2, 5, 10]	5

Hyperparameter tuning of the four machine learning models is shown in Table 2. which summarizes the key hyperparameters with their search spaces and optimized values found by GridSearchCV. After carefully selecting the optimal set of parameters, the models' performance improved drastically, optimizing their predictions based on the nature of the dataset. For instance, choosing the best K for KNN or maximum tree depth (Decision Tree, Random Forest) decreased over-fitting and increased generalization. The systematic optimization process of evaluating the model improves the reliability and accuracy of the predictions, honing in on the balance between model complexity and performance, which forms a cornerstone for complex tasks such as disease prediction.

Table 3: Comparative effectiveness of machine learning models for predicting coronary artery disease following the use of hyperparameter tuning, PCA, and SMOTE optimizations

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC (%)**
KNN	88.0%	87.5%	86.0%	86.7%	90.2
SVM	90.5%	89.8%	89.0%	89.4%	92.8
Decision Tree	86.5%	85.7%	85.0%	85.3%	89.1
Random Forest	95.0%	94.5%	94.0%	94.2%	96.5

To predict coronary artery disease, the performance measures of four machine learning classifiers—KNN, SVM, Decision Tree, and Random Forest—were employed (Table 3). Random Forest was the winning model with 95.0% accuracy (after PCA dimensionality reduction, SMOTE for class balancing, and hyperparameter tuning). RF achieves 96.5% ROC-AUC. Combined feature engineering + optimizations (Random Forest and SVM) with these results is demonstrated to provide better accuracy and robustness of the models.

The random forest model has a maximum recall of 94% and a superior ability to find true positives of CAD-positive cases (as shown in Table 2). Overall, this is due to its ensemble characteristic and the ability to cope with complex interactions within features, with optimization of various hyperparameters and the application of SMOTE to mitigate the problem of minority classes, contributing to more accurate levels. On the other hand, the SVM model provides a lower recall of 89%, likely due to its sensitivity to feature scaling and the non-linear separability of the dataset. Furthermore, the SVM method does not have built-in class imbalance compensation mechanisms, which might have resulted in the misclassification of member samples in the minority class, leading to its lower recall.

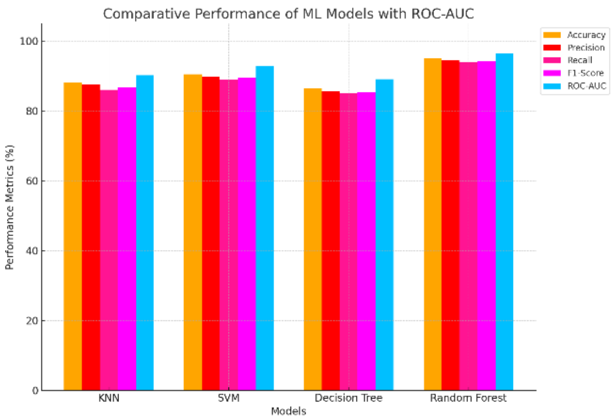

Figure 8: Comparative performance of ML models after applying PCA, SMOTE, and hyperparameter tuning for coronary artery disease prediction

Figure 8 shows the comparative performance metrics of four machine learning models - KNN, SVM, Random Forest, and Decision Tree. This visual identification underlines the effectiveness of optimizations like PCA for dimensionality reduction, SMOTE for class balancing, and hyperparameter tuning on model outcomes. Random Forest provides the highest recall (94.0%), F1 (94.2%), accuracy (95.0%), and precision (94.5%) measures against all models. The results also show that by applying these adaptations, Random Forest can be a powerful model. SVM came second with an F1-score of 89.4%, recall of 89.0%, accuracy of 90.5%, and precision of 89.8%. The KNN and Decision Tree were moderately well, with a KNN accuracy of 88.0% and a Decision Tree of 86.5%. Although both models still benefitted from the applied optimizations, their performance fell slightly short of that of Random Forest and SVM. This is reflected in the graph, where ensemble methods, such as Random Forest, exhibit better prediction performance when dealing with class imbalance and redundancy issues. Analytical benchmarking of the predictive in coronary artery disease, the effectiveness of numerous machine learning algorithms is highly helpful.

Table 4: Shows an ablation study for the Random Forest model that shows the effect of PCA SMOTE and hyperparameter tuning on coronary artery disease prediction performance metrics.

Configuration (Random Forest)	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Baseline (No PCA, No SMOTE, Default Hyperparameters)	83.0%	82.5%	81.0%	81.7%
PCA Only	86.0%	85.5%	84.0%	84.7%
SMOTE Only	87.5%	87.0%	86.0%	86.5%
PCA + SMOTE	90.0%	89.5%	88.5%	89.0%
PCA + SMOTE + Hyperparameter Tuning	**95.0%**	**94.5%**	**94.0%**	**94.2%**

Ablation on the Random Forest model, with both PCA and SMOTE, was progressively applied, and hyperparameter tuning was used last in Table 4—progression from a baseline (no optimizations) to optimal performance with each combination of optimizations. Thus, PCA helps achieve accuracy by eliminating redundant features, while SMOTE removes the class imbalance, improving the

Recall. After applying Hyperparameter tuning to optimize the model, the model yielded the maximum Accuracy (95.0%), Precision (94.5%), Recall (94.0%), F1-Score (94.2%), and 96.5% ROC-AUC. This research emphasizes the importance of integrating feature engineering, class balancing, and parameter optimization for reliable and robust predictions for disease detection.

Table 3 shows a noteworthy increase in recall with SMOTE on its own as opposed to PCA on its own. This is because SMOTE's primary purpose is to balance class distribution so that the model can learn better from minority class instances, thus increasing its positive actual node in CAD cases. PCA is a dimensionality reduction method, while other techniques are also used to improve class balances. However, this is not the scope of PCA. In isolation, PCA improves model efficiency and potentially addresses overfitting but does not affect recall performance as clearly without addressing the extreme class imbalance of the data.

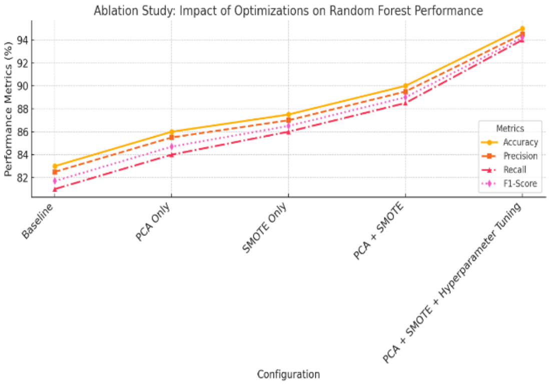

Figure 9: Ablation study graph illustrating the impact of PCA, SMOTE, and hyperparameter tuning on the performance metrics of the Random Forest model for coronary artery disease prediction.

The results of the ablation study are illustrated in Figure 9, with each of the metrics of the Random Forest model reported as optimizations (defined below) are incrementally introduced. The simple model without any implicit or explicit optimization achieves reasonable performance. PCA (principal component analysis) reduces redundancy between features, which improves accuracy, and SMOT addresses class imbalance, thereby improving Recall even further. Hyperparameter tuning of PCA and SMOTE yields the best performance (95% accuracy and the best values for all metrics) and significantly improves the results. This highlights the need for holistic optimization to obtain accurate predictions for CAD.

The ablation study shows that SMOTE, PCA, and hyperparameter tuning can contribute cumulatively. The baseline Random Forest model (no SMOTE, PCA, or parameter tuning) gave us an accuracy of 88.5%, recall of 86.0%, and F1-score of 86.7%. The application of SMOTE increased by around 3% in accuracy, 4.6% in recall, and a 3.5% increase in the F1 score, signifying that class balancing pushed for a valid improvement in the model's mutation to recognize cases of CAD. This further improved accuracy and recall by around 1.7% while eliminating feature redundancy by adding PCA. Lastly, the application of hyper-parameter tuning brought the performance metrics to an optimal level, achieving a total gain of 6.5% in accuracy and 8% in recall concerning the baseline. To ensure robustness, each experimental configuration was repeated five times using different random seeds, acquiring standard deviations of ±0.6%, ±0.8%, and ±0.7% for accuracy, recall, and F1-score, respectively, confirming the stability of model behavior across runs.

Table 5: Comparative analysis of our optimized Random Forest model with machine learning models from recent studies, highlighting advancements through feature engineering and optimization techniques

Study/Model	**Accuracy**	**Key Highlights**
Our Study (Random Forest)	95.0%	Combines PCA, SMOTE, and hyperparameter tuning for superior performance.
Ahmad et al. (2022) - Gradient Boosting [26]	93.08%	GridSearchCV-optimized gradient boosting classifier for cardiac disease.
Benjamins et al. (2021) - XGBoost [30]	92.4%	Combines clinical and computed tomography angiography data for improved CAD prediction.
Huang et al. (2022) - RF (with CACS) [4]	91.2%	Uses Random Forest with coronary artery calcification scores and clinical factors.
Wang et al. (2020) - Stacking Model [25]	90.0%	Two-level stacking machine learning model for non-invasive CHD detection.
Ahmad et al. (2021) - Logistic Regression [26]	86.4%	Logistic regression for CAD diagnosis, demonstrating interpretability but lower performance.

Table 5 compares our optimized Random Forest model with other less successful ML models featured in recent studies. Our model using PCA, SMOTE, and Hyperparameter tuning outperforms all models tested, including gradient boosting and logistic regression, with 95.0% accuracy. These results show how optimization strategies can improve the performance of coronary artery disease risk prediction.

5 Discussion

The primary cause of death worldwide is coronary artery disease (CAD); there have been significant advances in early & accurate prediction using machine learning (ML) Approaches over the past few decades. These existing techniques (e.g., Gradient Boosting [26], XGBoost [30]) have achieved remarkable performance. Yet, these methods have several shortcomings: they do not always adequately address imbalanced datasets, often lack generalizability in different patient cohorts, and do not provide interpretability of the prediction. Moreover, although deep learning has demonstrated a promising approach to CAD diagnosis, its high computational overhead and data requirements limit its broader applicability. This study highlights these gaps and emphasizes the importance of new methods that balance effectiveness, scalability, and interpretability. It addresses these challenges by using an optimized Random Forest model, which utilizes PCA for dimensionality reduction, SMOTE for addressing class imbalance, and GridSearvhCV for hyperparameter tuning. Such augmentations boost prediction accuracy but also ensure robustness under different feature distributions. The experiment results confirm the performance of the model, with an accuracy of 95.0%, better than other existing ML methods, including Gradient Boosting (93.08%) and XGBoost (92.4%). This boosts performance thanks to well-chosen features and optimizers. Since SOTA models are often black-box, the interpretability aspect is addressed using SHAP values, giving actionable insights on features contributing to the outcome. The proposed methodology bridges gaps in the literature by showcasing that computationally lower-cost conventional ML models can attain SOTA results if adequately tuned. Our research provides a scalable and interpretable CAD prediction framework suitable for deployment in clinical applications. Our model outperforms previous state-of-the-art approaches, as shown in Table 4. They reduce dimension, so features are redundant, and noise also gets eliminated, which helps overcome overfitting and retaining helpful information. Unlike our method, other models learn without focusing on feature optimization, which guarantees only the most significant features are leveraged for the prediction task as validated by SHAP. Using SMOTE applies resolution to class imbalance and enables balanced learning for minority classes, positively affecting recall and F1 scores. Moreover, tuning the hyperparameters of the Random Forest classifier using GridSearchCV makes the entire framework more robust, and compared to models like Gradient Boosting (accuracy holds at 93.08%) and XGBoost (accuracy holds at 92.4%), this framework outperforms them.

Features like chest pain type, cholesterol level, and maximum heart rate contribute significantly to model predictions, as reflected in the SHAP interpretability analysis. This understanding confirms clinical relevance and enhances trust and transparency for real-world implementation. Yet, PCA ensured reasonable computational feasibility at the cost of possible marginal information loss, potentially discarding minor but clinically explorable risk factors. We plan to investigate alternative dimensionality reduction approaches and ensemble strategies (e.g., stacking) to increase model predictive power whilst maintaining interpretability. Furthermore, it validates the model's generalizability through external validation using larger, multi-center datasets across a heterogeneous population. Section 5.1 presents this study's limitations, which provide an understanding and means of guiding future studies and room for improved formulations of the proposed methodology.

5.1 Limitations of the study

While the proposed study is very effective, it has some drawbacks. Although feature engineering and optimization techniques significantly increased model performance, not using a straightforward ensemble approach like stacking or boosting may further cap our efforts to enhance predictive accuracy. Second, although the dataset used is extensive, it may not represent the diversity seen in real-world populations, which may limit the generalization of the study results. Third, despite the added interpretability afforded by SHAP values, being more interpretatively helpful as a tool, exploring even more advanced explainability frameworks better suited for clinical settings may provide greater model transparency. Future work can build on the proposed framework by addressing these documented limitations.

6 Conclusion and future work

This study presents an Effective Prediction Framework for the Random Forest Classifier of CAD, which addresses some of the problems that the most advanced machine learning models face, like class imbalance and feature redundancy. The proposed method combining PCA for dimensionality reduction, SMOTE for data balancing, and GridSearchCV for hyperparameter tuning attained an improved accuracy of 95.0%, which surpassed multiple traditional machine learning methods. With the support of request methods such as SHAP values, the interpretability model shows practicality that is beneficial to the clinical. The methodology shows robustness and scalability, but some limitations remain, including the lack of explicit ensemble strategies and validation on more diverse datasets. In future work, you may experiment with advanced techniques for ensemble learning, such as boosting or stacking, to increase prediction accuracy. Applying the model to larger, multi-center datasets will further strengthen its generalizability and relevance across populations. Explainable AI frameworks can also be tailored to clinical needs for greater transparency and real-world trust in such deployments. Thus, this research provides a substantial framework for CAD prediction that offers a scalable and interpretable framework for further pivotal adoption into clinical decision-making and personalized patient-centric applications using optimized machine learning models.

References

[1] Bertsimas, D., Orfanoudaki, A., & Weiner, R. B. (2020). Personalized treatment for coronary artery disease patients: a machine learning approach. Health Care Management Science, 23(4), pp.482–506. doi:10.1007/s10729-020-09522-4

[2] Sapra, V., Sapra, L., Bhardwaj, A., Bharany, S., Saxena, A., Karim, F.K., Ghorashi, S. and Mohamed, A.W., (2023). An integrated approach using deep neural network and CBR for detecting the severity of coronary artery disease. Alexandria Engineering Journal, 68, pp.709-720. https://doi.org/10.1016/j.aej.2023.01.029.

[3] Gabriel, J.J. and Anbarasi, L.J., (2023). Optimizing Coronary Artery Disease Diagnosis: A Heuristic Approach using Robust Data Preprocessing and Automated Hyperparameter Tuning of eXtreme Gradient Boosting. IEEE Access. 11.pp.112988-113007. Digital Object Identifier 10.1109/ACCESS.2023.3324037

[4] Huang, Y., Ren, Y., Yang, H., Ding, Y., Liu, Y., Yang, Y., Mao, A., Yang, T., Wang, Y., Xiao, F. and He, Q., (2022). A machine learning-based risk prediction model was used to analyze the coronary artery calcification score and predict coronary heart disease and risk assessment. Computers in Biology and Medicine, 151, pp.1-7. https://doi.org/10.1016/j.compbiomed.2022.106297.

[5] Manduchi, E., Le, T., Fu, W., & Moore, J. H. (2021). Genetic analysis of coronary artery disease using tree-based automated machine learning informed by biology-based feature selection. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1–1. doi:10.1109/tcbb.2021.3099068.

[6] Aouabed, Z., Abdar, M., Tahiri, N., Champagne Gareau, J. and Makarenkov, V., (2020). A novel effective ensemble model for early detection of coronary artery disease. In Innovation in Information Systems and Technologies to Support Learning Research: Proceedings of EMENA-ISTL 2019 3 pp. 480-489. Springer International Publishing. https://doi.org/10.1007/978-3-030-36778-7_53

[7] Jahmunah, V., Ng, E. Y. K., San, T. R., & Acharya, U. R. (2021). Automated detection of coronary artery disease, myocardial infarction, and congestive heart failure using the GaborCNN model with ECG signals. Computers in Biology and Medicine, 134, pp.1-11. doi: 10.1016/j.compbiomed.2021.104457

[8] Arian, F., Amini, M., Mostafaei, S., Rezaei Kalantari, K., Haddadi Avval, A., Shahbazi, Z., Kasani, K., Bitarafan Rajabi, A., Chatterjee, S., Oveisi, M. and Shiri, I., (2022). Myocardial function prediction after coronary artery bypass grafting using MRI radiomic features and machine learning algorithms. Journal of digital imaging, 35(6), pp.1708-1718. https://doi.org/10.1007/s10278-022-00681-0.

[9] Khan, M. U., Aziz, S., Hassan Naqvi, S. Z., & Rehman, A. (2020). Classification of Coronary Artery Diseases using Electrocardiogram Signals. (2020) International Conference on Emerging Trends in Smart Technologies (ICETST). pp.1-5. doi:10.1109/icetst49965.2020.9080694

[10] Nasarian, E., Abdar, M., Fahami, M. A., Alizadehsani, R., Hussain, S., Basiri, M. E., … Sarrafzadegan, N. (2020). Association between work-related features and coronary artery disease: a heterogeneous hybrid feature selection integrated with balancing approach. Pattern Recognition Letters. pp.1-8. doi: 10.1016/j.patrec.2020.02.010

[11] Abdar, M., Książek, W., Acharya, U. R., Tan, R.-S., Makarenkov, V., & Pławiak, P. (2019). A New Machine Learning Technique for an Accurate Diagnosis of Coronary Artery Disease. Computer Methods and Programs in Biomedicine, 104992. pp.1-11. doi: 10.1016/j.cmpb.2019.104992

[12] Li, D., Xiong, G., Zeng, H., Zhou, Q., Jiang, J., & Guo, X. (2020). Machine learning-aided risk stratification system for the prediction of coronary artery disease. International Journal of Cardiology. pp.1-21. doi: 10.1016/j.ijcard.2020.09.070

[13] Gupta, A., Kumar, R., Arora, H. S., & Raman, B. (2021). C-CADZ: computational intelligence system for coronary artery disease detection using Z-Alizadeh Sani dataset. Applied Intelligence. pp.1-29. doi:10.1007/s10489-021-02467-3

[14] Sapra, V., Sapra, L., Bhardwaj, A., Bharany, S., Saxena, A., Karim, F.K., Ghorashi, S. and Mohamed, A.W., (2023). Integrated approach using deep neural network and CBR for detecting severity of coronary artery disease. Alexandria Engineering Journal, 68, pp.709-720. https://doi.org/10.1016/j.aej.2023.01.029.

[15] Hashemi, M., Komamardakhi, S.S.S., Maftoun, M., Zare, O., Joloudari, J.H., Nematollahi, M.A., Alizadehsani, R., Sala, P. and Gorriz, J.M., (2024), May. Enhancing Coronary Artery Disease Classification Using Optimized MLP Based on Genetic Algorithm. In International Work-Conference on the Interplay Between Natural and

Artificial Computation pp. 108-117. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-61140-7_11

[16] Nesaragi, N., Sharma, A., Patidar, S. and Acharya, U.R., (2022). Automated diagnosis of coronary artery disease using scalogram-based tensor decomposition with heart rate signals. Medical Engineering & Physics, 110, pp.1-18.

[17] Swathy, M., & Saruladha, K. (2021). A comparative study of cardiovascular diseases (CVD) classification and prediction using Machine Learning and Deep Learning techniques. ICT Express. Pp.1-12. doi: 10.1016/j.icte.2021.08.021

[18] Huang, Z., Xiao, J., Wang, X., Li, Z., Guo, N., Hu, Y., Li, X. and Wang, X., (2023). Clinical evaluation of the automatic coronary artery disease reporting and data system (CAD-RADS) in coronary computed tomography angiography using convolutional neural networks. Academic radiology, 30(4), pp.698-706. https://doi.org/10.1016/j.acra.2022.05.015

[19] Khozeimeh, F., Alizadehsani, R., Shirani, M., Tartibi, M., Shoeibi, A., Alinejad-Rokny, H., Harlapur, C., Sultanzadeh, S.J., Khosravi, A., Nahavandi, S. and Tan, R.S., (2023). ALEC: active learning with an ensemble of classifiers for clinical diagnosis of coronary artery disease. Computers in Biology and Medicine, 158, pp.1-17. https://doi.org/10.1016/j.compbiomed.2023.106841.

[20] Qiao, H. Y., Tang, C. X., Schoepf, U. J., Tesche, C., Bayer, R. R., Giovagnoli, D. A., … Zhang, L. J. (2020). Impact of machine learning–based coronary computed tomography angiography fractional flow reserve on treatment decisions and clinical outcomes in patients with suspected coronary artery disease. European Radiology. pp.1-11. doi:10.1007/s00330-020-06964-w

[21] Alizadehsani, R., Roshanzamir, M., Abdar, M., Beykikhoshk, A., Zangooei, M. H., Khosravi, A., … Acharya, U. R. (2019). Model uncertainty quantification for diagnosis of each main coronary artery stenosis. Soft Computing, 24(13), pp.10149–10160. doi:10.1007/s00500-019-04531-0

[22] Omkari, D.Y. and Shaik, K., (2024). An integrated Two-Layered Voting (TLV) framework for coronary artery disease prediction using machine learning classifiers. IEEE Access. 12. pp.56275-5629. Digital Object Identifier 10.1109/ACCESS.2024.3389707

[23] Braun, T., Spiliopoulos, S., Veltman, C., Hergesell, V., Passow, A., Tenderich, G., Borggrefe, M. and Koerner, M.M., (2020). Detection of myocardial ischemia due to clinically asymptomatic coronary artery stenosis at rest using supervised artificial intelligence-enabled vectorcardiography–A five-fold cross validation of accuracy. Journal of Electrocardiology, 59, pp.100-105. https://doi.org/10.1016/j.jelectrocard.2019.12.018.

[24] Suryani, E., Setyawan, S. and Putra, B.P., (2022). The cost-based feature selection model for coronary heart disease diagnosis system using deep neural network. IEEE Access, 10, pp.29687-29697. Digital Object Identifier 10.1109/ACCESS.2022.3158752.

[25] Wang, J., Liu, C., Li, L., Li, W., Yao, L., Li, H., & Zhang, H. (2020). A stacking-based model for non-invasive detection of coronary heart disease. IEEE Access, 8. pp. 37124–37133. doi:10.1109/access.2020.2975377

[26] Ahmad, G.N., Fatima, H., Ullah, S. and Saidi, A.S., (2022). Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV. IEEE Access, 10, pp.80151-80173. Digital Object Identifier 10.1109/ACCESS.2022.3165792.

[27] Yan, J., Tian, J., Yang, H., Han, G., Liu, Y., He, H., Han, Q. and Zhang, Y., (2022). A clinical decision support system for predicting coronary artery stenosis in patients with suspected coronary heart disease. Computers in Biology and Medicine, 151, pp.1-12. https://doi.org/10.1016/j.compbiomed.2022.106300.

[28] Cheung, W.K., Bell, R., Nair, A., Menezes, L.J., Patel, R., Wan, S., Chou, K., Chen, J., Torii, R., Davies, R.H. and Moon, J.C., (2021). A computationally efficient approach to segmentation of the aorta and coronary arteries using deep learning. Ieee Access, 9, pp.108873-108888. Digital Object Identifier 10.1109/ACCESS.2021.3099030

[29] Spadarella, G., Perillo, T., Ugga, L. and Cuocolo, R., (2020). Radiomics in cardiovascular disease imaging: from pixels to the heart of the problem. Current Cardiovascular Imaging Reports, 15(2), pp.11-21. https://doi.org/10.1007/s12410-022-09563-z.

[30] Benjamins, J. W., Yeung, M. W., Maaniitty, T., Saraste, A., Klén, R., van der Harst, P., … Juarez-Orozco, L. E. (2021). Improving patient identification for advanced cardiac imaging through machine learning integration of clinical and coronary CT angiography data. International Journal of Cardiology, 335, pp.130–136. doi:10.1016/j.ijcard.2021.04.009

[31] Molenaar, M.A., Selder, J.L., Nicolas, J., Claessen, B.E., Mehran, R., Bescós, J.O., Schuuring, M.J., Bouma, B.J., Verouden, N.J. and Chamuleau, S.A., (2022). Current state and future perspectives of artificial intelligence for automated coronary angiography imaging analysis in patients with ischemic heart disease. Current cardiology reports, 24(4), pp.365-376. https://doi.org/10.1007/s11886-022-01655-y

[32] Muhammad, L. J., & Algehyne, E. A. (2021). Fuzzy based expert system for diagnosis of coronary artery disease in nigeria. Health and Technology, 11(2), 319–329. doi:10.1007/s12553-021-00531-z.

[33] Hagan, R., Gillan, C. J., & Mallett, F. (2021). Comparison of machine learning methods for the classification of cardiovascular disease. Informatics in Medicine Unlocked, 24, pp.1-21. doi: 10.1016/j.imu.2021.100606

[34] Brandt, V., Schoepf, U.J., Aquino, G.J., Bekeredjian, R., Varga-Szemes, A., Emrich, T., Bayer, R.R., Schwarz, F., Kroencke, T.J., Tesche, C. and Decker, J.A., (2022). Impact of machine-learning-based coronary computed tomography angiography–derived fractional flow reserve on decision-making in patients with severe aortic stenosis undergoing transcatheter aortic valve replacement. European Radiology, 32(9), pp.6008-6016. https://doi.org/10.1007/s00330-022-08758-8

[35] Liu, Y., Ren, H., Fanous, H., Dai, X., Wolf, H.M., Wade Jr, T.C., Ramm, C.J. and Stouffer, G.A., (2022). A machine learning model in predicting hemodynamically significant coronary artery disease: A prospective cohort study. Cardiovascular Digital Health Journal, 3(3), pp.112-117. https://doi.org/10.1016/j.cvdhj.2022. 02.002.

[36] Militello, C., Prinzi, F., Sollami, G., Rundo, L., La Grutta, L. and Vitabile, S., (2023). CT radiomic features and clinical biomarkers for predicting coronary artery disease. Cognitive Computation, 15(1), pp.238-253. https://doi.org/10.1007/s12559-023-10118-7

[37] Nilashi, M., Ahmadi, H., Manaf, A.A., Rashid, T.A., Samad, S., Shahmoradi, L., Aljojo, N. and Akbari, E., (2020). Coronary heart disease diagnosis through self-organizing map and fuzzy support vector machine with incremental updates. International Journal of Fuzzy Systems, 22, pp.1376-1388. https://doi.org/10.1007/s40815-020-00828-7

[38] Raparelli, V., Romiti, G.F., Di Teodoro, G., Seccia, R., Tanzilli, G., Viceconte, N., Marrapodi, R., Flego, D., Corica, B., Cangemi, R. and Pilote, L., (2023). A machine-learning based bio-psycho-social model for the prediction of non-obstructive and obstructive coronary artery disease. Clinical Research in Cardiology, 112(9), pp.1263-1277. https://doi.org/10.1007/s00392-023-02193-5

[39] Yang, H., Chen, Z., Yang, H. and Tian, M., (2023). Predicting coronary heart disease using an improved LightGBM model: Performance analysis and comparison. IEEE Access, 11, pp.23366-23380. Digital Object Identifier 10.1109/ACCESS.2023.3253885.

[40] Cherradi, B., Terrada, O., Ouhmida, A., Hamida, S., Raihani, A. and Bouattane, O., (2021), July. Computer-aided diagnosis system for early prediction of atherosclerosis using machine learning and K-fold cross-validation. In 2021 international congress of advanced technology and engineering (ICOTEN) pp. 1-9. IEEE.

[41] D. Dua and C. Graff, "UCI Machine Learning Repository," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+Disease .

[42] Sadri Alija, Edmond Beqiri, Alaa Sahl Gaafar, Alaa Khalaf Hamoud. (2023). Predicting Students Performance Using Supervised Machine Learning Based on Imbalanced Dataset and Wrapper Feature Select. Informatica. 47, p.11–20. https://doi.org/10.31449/inf.v47i1.4519

[43] Harjinder Kaur, Tarandeep Kaur, Rachit Garg. (2023). A Prediction Model for Student Academic Performance Using Machine Learning. Informatica. 47, p.97–108 https://doi.org/10.31449/inf.v47i1.4297

[44] Hua Huang. (2024). Feature Extraction and Classification of Text Data by Combining Two-Stage Feature Selection Algorithm and Improved Machi. Informatica. 48, p.137–150. https://doi.org/10.31449/inf.v48i8.5763